Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

June 2026. Vol. 5. No. 3; e-ISSN: 2808-4519

Implementation of the Random Forest Algorithm for Loan Eligibility Prediction and Feature Analysis Based on Financial Data

Angel[1*], Joni[2], Herman[3]

[1,2,3]*STMIK TIME, Medan, Indonesia*

angellim2419@gmail.com[1*], joni.hgw@gmail.com[2], hrmn_ang@yahoo.com[3]

Abstract

The advancement of information technology has led to an increasing demand for loan access, both through banking institutions and online lending platforms. However, the process of evaluating loan eligibility, which is still carried out manually or semi-manually, is prone to human error and decision-making bias, ultimately increasing the risk of loan defaults. This study aims to implement the Random Forest algorithm to predict loan eligibility based on financial data, as well as to evaluate its accuracy. The dataset used in this study is loan_approval_dataset.csv, which is downloaded from Kaggle, utilizing 11 input features. The system is developed as a web-based application using Laravel as the main frontend and backend framework, while Flask is used as a backend API for executing the machine learning processes. The testing results show that Random Forest model achieves an accuracy of 98.44%, with a precision of 98.14%, recall of 99.37%, and an F1-score of 98.75%. Furthermore, the cibil score feature is identified as the most influential factor in the prediction process, contributing 80.65% to the model's outcome. These findings indicate that the Random Forest algorithm is highly effective for use in a loan eligibility prediction system, as it provides fast, objective, and highly accurate results.

Keywords: *CIBIL Score; Data Mining; Loan Eligibility Prediction; Machine Learning; Random Forest*

1. Introduction

The development of information and communication technology has significantly impacted various aspects of society, particularly in the economic sector. This impact was further amplified by the COVID-19 pandemic, which caused a sharp decline in economic indicators, weakened purchasing power, and increased public anxiety over job loss, supply chain uncertainty, and income reduction [1]. As economic stability deteriorated, access to financing, such as loans, has become greater in order to support daily living needs.

Loans are available in various forms, including bank credit and online lending (fintech). The word of "credit" originates from the Latin "credere," meaning trust, and refers to the provision of money or bills based on an agreement between a bank and a borrower to be repaid within a certain period [2]. In contrast, online loans offer simplier requirements and faster processing [3]. However, both forms of lending frequently misjudge the eligibility of loan applicants. Manual or semi-manual evaluation processes are prone to human error and decision bias, resulting in high default rates and rejection of qualified applicants.

To address these limitations, this study proposes the implementation of a machine learning approach, specifically the Random Forest (RF) algorithm, for automated loan eligibility prediction. Random Forest is an ensemble method consisting of multiple decision trees that classify data into classes and can improve accuracy on large training datasets [4]. Its key advantages include high accuracy, robustness to imbalanced data, and resistance to overfitting [5].

This study contributes a fully functional web-based prediction system integrating Laravel for the main application and Flask as the machine learning API backend. The system not only predicts loan eligibility but also provides feature importance analysis, offering interpretable insights for financial institutions.

2. Literature Review

2.1. Loan and credit

Credit is the provision of funds based on an agreement between a bank and a borrower, obligating the borrower to repay the debt within a specified period with interest. The core principle of credit is trust, supported by elements such as agreement, loan term, risk, and profit margin. Credit quality is classified under Bank Indonesia Regulation No. 9/6/PBI/2007 into Performing Loan (no arrears beyond 90 days) and Non-Performing Loan (NPL), the latter representing serious repayment issues with direct impact on bank financial health [6]. .

2.2. Online lending (fintech lending)

Online lending (Fintech Lending) uses digital technology to provide borrowing services without face-to-face interaction. While regulated providers registered with the Financial Services Authority (OJK) apply proper verification procedures, illegal operators often use deceptive incentives with excessively high interest rates. Unresolved debts can lead to aggressive debt collection, harassment, data abuse, and severe psychological consequences for borrowers [7].

2.3. Related work

Several prior studies have applied Random Forest in various prediction domains, as summarized in Table 1.

Table 1: Summary of related work on Random Forest prediction

Author	Topic	Algorithm	Accuracy
Azmi & Voutama [8]	Bank customer churn prediction	Random Forest	Higher than Decision Tree
Azzahra et al. [9]	Stunting data prediction	Random Forest + Cross Validation	77.55%
Husen et al. [10]	Forest fire prediction	Random Forest Classifier	100%
Maulidah et al. [11]	Water quality prediction	Random Forest	88.33%
Tamba & E. [12]	Heart failure disease prediction	Random Forest	82.60%

These studies consistently demonstrate that Random Forest achieves strong performance across diverse prediction tasks, providing a sound basis for its application to loan eligibility prediction in the present study.

3. Research Methodology

3.1. Dataset

The study utilises the publicly available *loan_approval_dataset.csv* obtained from Kaggle [13], containing 4,269 borrower records. Following removal of the non-predictive *loan_id* identifier, eleven input features remain. Table 2 summarises the input parameters.

Table 2: Input Feature Summary

No.	Feature	Type	Description
1	*no_of_dependents*	Numeric	Number of dependents (0–5+)
2	*education*	Categorical	Education level (Graduate / Not Graduate)
3	*self_employed*	Categorical	Employment status (Yes / No)
4	*income_annum*	Numeric	Annual income (IDR)
5	*loan_amount*	Numeric	Requested loan amount (IDR)
6	*loan_term*	Numeric	Loan duration (months)
7	*cibil_score*	Numeric	Credit score (300–900)
8	*residential_assets_value*	Numeric	Value of residential assets (IDR)

9	*commercial_assets_value*	Numeric	Value of commercial assets (IDR)
10	*luxury_assets_value*	Numeric	Value of luxury assets (IDR)
11	*bank_assets_value*	Numeric	Value of bank assets (IDR)

3.2. Data preprocessing

The preprocessing steps applied were as follows:

(1) Removal of non-predictive columns (loan_id). (2) Whitespace trimming from all text fields. (3) Standardization of loan_status labels to lowercase: "approved" (encoded as 1) and "rejected" (encoded as 0). (4) Removal of rows with empty loan_status labels. (5) Imputation of missing values using median for numerical features and mode for categorical features. After preprocessing, 4,269 valid records were retained.

3.3. Data transformation

Features were separated into input matrix X (11 features) and target vector y (loan_status). Categorical variables (education, self_employed) were encoded using One-Hot Encoding. Numerical features were standardized using Standard Scaler (mean = 0, standard deviation = 1) to ensure uniform feature scales.

3.4. Data splitting

The dataset was split into training (70%) and testing (30%) sets: training set = 2,988 records; test set = 1,281 records. Splitting was performed using stratified random sampling to preserve class proportions.

3.5. Model training

A *Random Forest Classifier* was trained using the following hyperparameters: n_estimators = 310, max_depth = 15, random_state = 0. The model was trained on (X_train, y_train) to learn the relationship between borrower financial features and loan eligibility outcome.

3.6 Model evaluation

Model performance was assessed on the test set using the confusion matrix and four derived metrics:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \times 100\%$$
$$Precision = TP / (TP + FP) \times 100\%$$
$$Recall = TP / (TP + FN) \times 100\%$$
$$F1 - Score = 2 \times (Precision \times Recall) / (Precision + Recall)$$

3.7. Feature importance analysis

Feature importance was computed based on the mean impurity decrease (Gini importance) contributed by each feature across all 310 decision trees in the ensemble. Features with higher cumulative impurity reduction were ranked as more important.

3.8. System architecture

The web application was developed using Laravel as the main frontend and backend framework, handling user authentication (session-based), dashboard, data management, prediction history, and reporting. Flask was deployed as a dedicated backend API serving the /predict endpoint, which receives borrower data in JSON format, performs preprocessing and prediction using the pre-trained model (.pkl), and returns the eligibility result along with a confidence score. Results are stored in a MySQL database across three tables: USERS, Loan_Data, and Prediction_History.

4. Results and discussion

4.1 Model performance

The Random Forest model evaluated on the 1,281-record test set achieved the performance metrics shown in Table 3.

Table 3: Model Evaluation Result

Metric	Value
Accuracy	98.44%

Precision	98.14%
Recall	99.37%
F1-Score	98.75%

The accuracy of 98.44% confirms that the model correctly classifies loan eligibility in nearly all test cases. The high precision (98.14%) indicates that most applicants predicted as eligible are genuinely creditworthy, minimizing erroneous approvals. The recall of 99.37% demonstrates that the model captures almost all truly eligible applicants, reducing wrongful rejections. The F1-score of 98.75% reflects an excellent balance between precision and recall, validating the model's suitability for practical deployment in lending institutions.

4.2. Feature importance analysis

Feature importance scores extracted from the Random Forest model are presented in Table 4.

Table 1.2 Feature importance ranking

Rank	Feature	Importance (%)
1	CIBIL Score	80.65%
2	Loan Term	6.4%
3	Loan Amount	3.05%
4	Income Annum	1.93%
5	Residential Assets Value	1.82%
6	Commercial Assets Value	1.8%
7	Bank Assets Value	1.62%
8	Luxury Assets Value	1.47%
9	No. of Dependents	0.77%
10	Education	0.26%
11	Self Employed	0.23%

CIBIL score dominates the prediction with a contribution of 80.65%, confirming that credit history is the most critical determinant of loan eligibility. This finding aligns with established credit risk theory, which positions credit scores as the primary indicator of a borrower's repayment capability. Loan term (6.40%) and loan amount (3.05%) rank second and third, reflecting the influence of loan structure on default risk. Asset values collectively contribute approximately 6%, while demographic features (education, self-employment status) have minimal influence (<0.5%), suggesting that financial behavior outweighs demographic factors in predicting creditworthiness.

4.3. System implementation

The web-based system was successfully implemented with the following modules.

Figure 4.1 shows the login page. This page serves as the user's entry point into the system. It contains two input fields, namely username and password, as well as a login button to access the system, i.e., the dashboard.

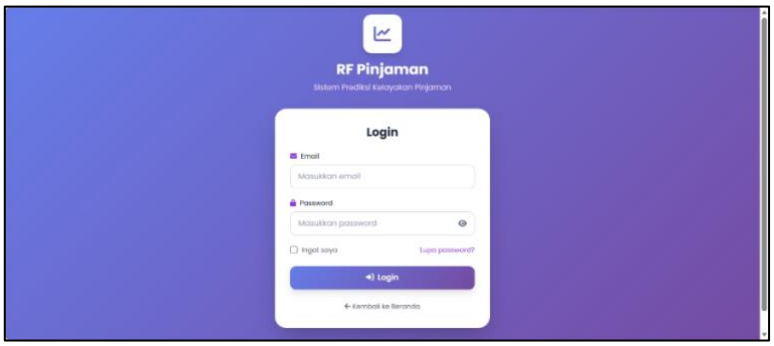

Fig.1: Login page

Figure 4.2 shows the dashboard view. This page displays the user currently logged in, the total data, the approval rate containing the approval percentage and the number of predictions deemed eligible and ineligible, the average CIBIL score, the average loan amount, and the model performance in the form of a bar chart.

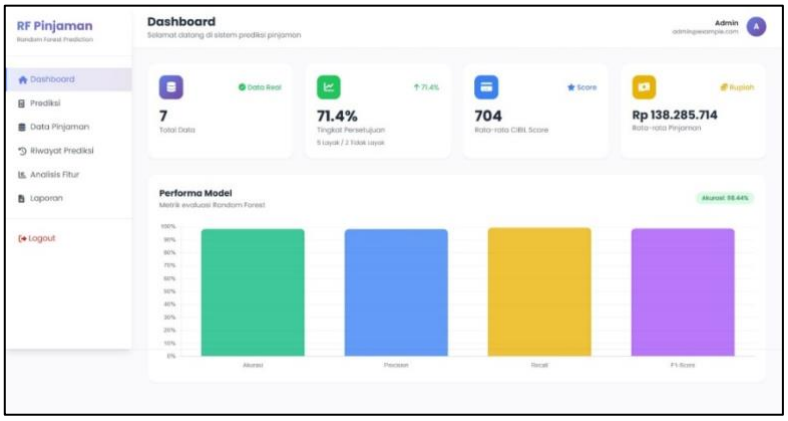

Fig.2: Dashboard page

Figure 4.3 shows the prediction page. The Prediction page displays a form used to predict loan eligibility using the Random Forest algorithm. On this page, the user must fill in the required data as input for the prediction model. After all data has been entered, the user can press the Process button to run the prediction process and obtain the loan-eligibility analysis result.

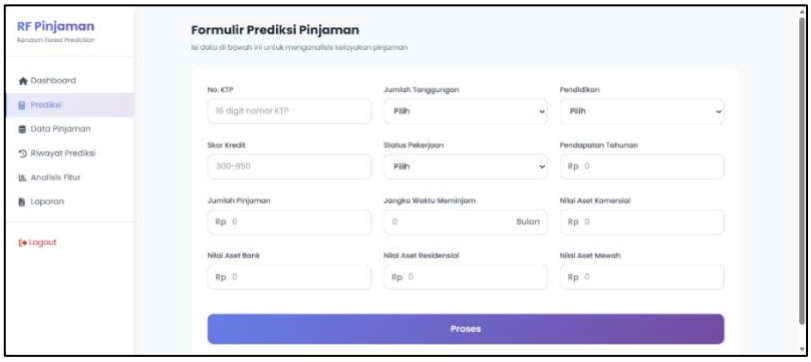

Fig.1: Prediction page

Figure 4.4 shows the prediction result page. This page displays the output of the prediction process that has been executed. At the top of the page, the system shows the prediction result status, namely “eligible” or “not eligible,” along with its probability percentage. The page also presents a summary of the data previously entered. On the right side, there is a donut chart visualization illustrating the prediction probability, as well as information on model performance such as accuracy and prediction time, plus follow-up recommendations. Users can also perform a new prediction, view the prediction history, and print the prediction result using the available buttons.

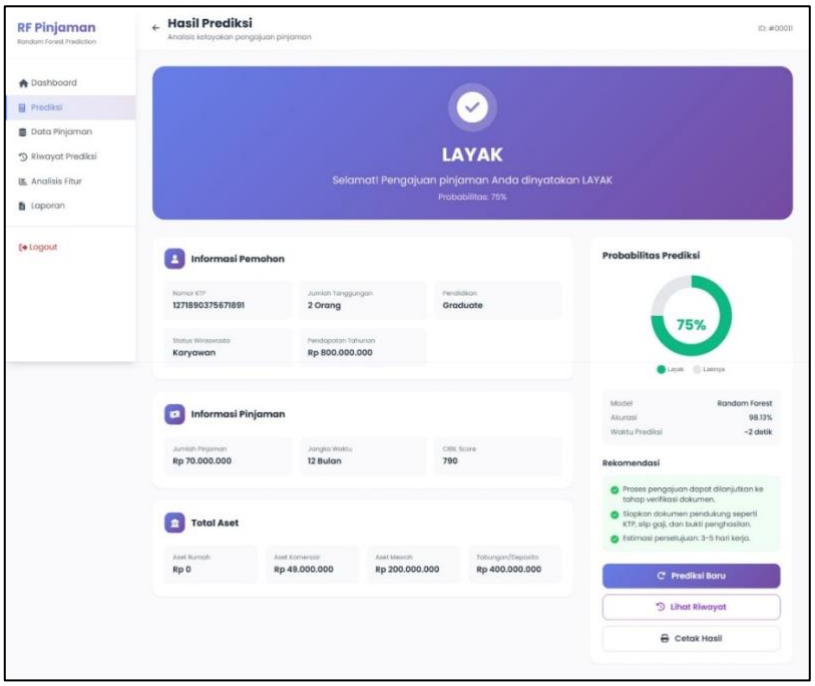

Figure 4.2 Prediction result page

Figure 4.5 shows the loan data page. The loan data page displays all applicant records stored in the system in a table format. The page is equipped with a search feature by ID number, filters by eligibility status, and an export button to download the data. Users can also add new records using the "Add Data" button and delete entries using the available icons.

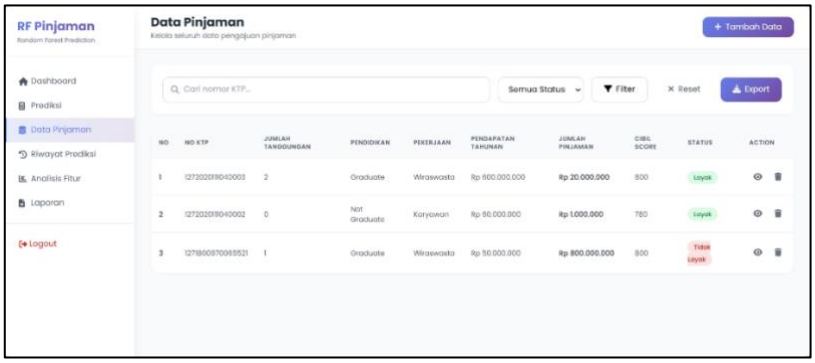

Fig.4: Loan Data Page

Figure 4.6 shows the prediction history page. The prediction history page displays all predictions made by the user. The data is presented in a table containing information such as number of dependents, education level, occupation, annual income, loan amount, CIBIL score, and the prediction result status. This page also includes a filter feature by prediction status and a "Make Prediction" button to perform a new prediction.

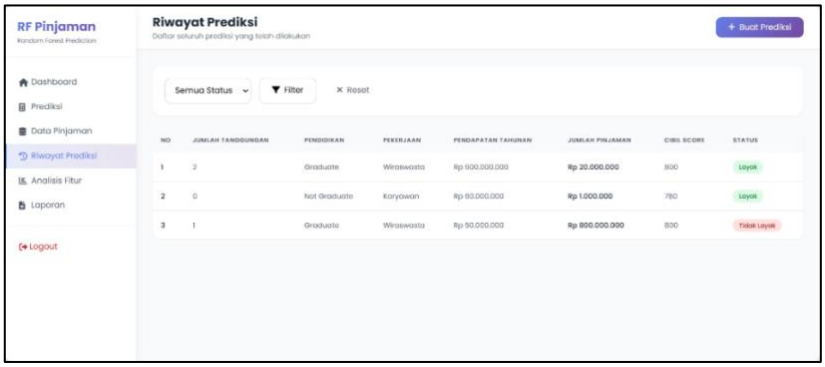

Fig.5: Prediction History Page

Figure 4.7 shows the feature analysis page. This page displays information on the importance level of each feature (feature importance) used by the Random Forest model in the prediction process. Each feature is presented together with its percentage contribution to the model in the form of a progress bar, arranged from the feature with the highest contribution

to the lowest. Based on the analysis results, the CIBIL score feature dominates with a contribution of 80.65%, followed by loan term at 6.4% and loan amount at 3.05%. The page also provides a “Retrain Model” button and an “Export Report” button, which can be used to download the feature analysis results.

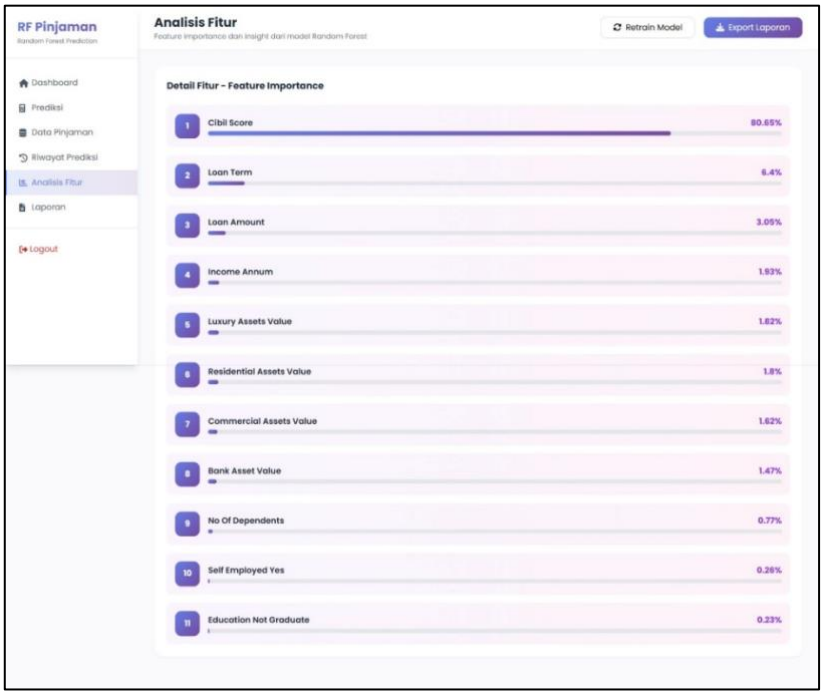

Fig.6: Feature Analysis page

Figure 4.8 shows the report page. The report page presents a summary of data and overall prediction model performance. At the top, there are general statistics covering the total data, the number of records with “eligible” status, and the number of records with “not eligible” status along with their percentages. In the model summary section, the model evaluation metrics include an accuracy of 98.44%, precision of 98.14%, recall of 99.37%, and an F1-score of 98.75%. In addition, this page also provides information on model configuration such as the number of estimators, max depth, and random state, as well as details of the dataset used in this study. Users can download the complete report using the “Export Report” button.

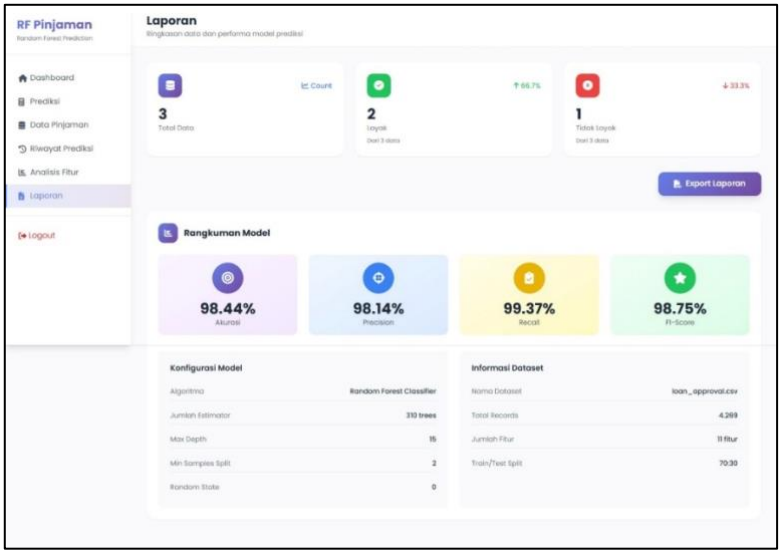

Fig.7: Reports Page

5. Conclusion

This study successfully implemented the Random Forest algorithm for loan eligibility prediction using financial data from the loan_approval_dataset (4,269 records, 11 features). The model achieved an accuracy of 98.44%, precision of 98.14%, recall of 99.37%, and F1-score of 98.75%, demonstrating superior classification performance suitable for real-world deployment in financial institutions.

CIBIL score was identified as the overwhelmingly dominant predictive feature (80.65%), followed by loan term (6.40%) and loan amount (3.05%). These findings provide actionable insights for lenders, confirming that credit history should remain the primary criterion in eligibility assessment.

The integrated web application that built with Laravel and Flask provides an end-to-end automated prediction system with interpretable feature analysis. Future work should explore multi-algorithm comparison (e.g., XGBoost, SVM), cross-validation evaluation strategies, expansion to mobile platforms, and integration of additional socioeconomic features to further improve prediction fairness and robustness.

Acknowledgement

The author would like to thank the thesis supervisors, faculty members, and all parties who supported this research. The dataset used in this study was sourced from Kaggle and made publicly available by A. Sharma [13].

References

[1] D. Anugrah, T. Tendiyanto, and S. Akhmaddhian, "Sosialisasi Bahaya Produk Pinjaman Online Ilegal bagi Masyarakat," *Empowerment*, vol. 4, no. 03, pp. 293–297, 2021, doi: 10.25134/empowerment.v4i03.5093.

[2] P. Permatasari, "Force Majeure Clausules Due To Covid-19 in Bank Credit Agreements," *Iblam Law Rev.*, vol. 1, no. 1, pp. 163–183, 2021, doi: 10.52249/ilr.v1i01.8.

[3] K. Kholidiah and T. Inayati, "Bijak Dalam Pengambilan Keputusan Pinjaman Online (Pinjol)," *JMM - J. Masy. Merdeka*, vol. 7, no. 1, p. 56, 2024, doi: 10.51213/jmm.v7i1.150.

[4] F. Diba, M. S. Lydia, and P. Sihombing, "Analisis Random Forest Menggunakan Principal Component Analysis Pada Data Berdimensi Tinggi," *Indones. J. Comput. Sci.*, vol. 12, no. 4, pp. 2152–2160, 2023, doi: 10.33022/ijcs.v12i4.3329.

[5] M. N. Raza, "Sistem Deteksi Berita Hoax Menggunakan Algoritma Naïve Bayes Dan Random Forest Pada Machine Learning," *Pondasi J. Appl. Sci. Eng.*, vol. 1, no. 2, pp. 43–57, 2024, [Online]. Available: https://journal.alshobar.or.id/index.php/pondasi/article/view/221

[6] W. Djuarni, "IMPLEMENTASI PRINSIP 5C DALAM MENENTUKAN KELAYAKAN PEMBERIAN KREDIT PADA NASABAH Wenny Djuarni 1 ," *Keuang. dan Perbank.*, vol. 02, no. 02, pp. 108–109, 2022.

[7] K. D. Sartika and D. Larasati, "Literature Review: Dampak Fenomena Pinjaman Online Ilegal di Indonesia," *Innov. J. Soc. Sci. Res.*, vol. 3, no. 6, pp. 2940–2948, 2023, [Online]. Available: https://j-innovative.org/index.php/Innovative/article/view/6517

[8] A. F. Azmi and A. Voutama, "Prediksi Churn Nasabah Bank Menggunakan Klasifikasi Random Forest Dan Decision Tree Dengan Evaluasi Confusion Matrix," *Komputa J. Ilm. Komput. dan Inform.*, vol. 13, no. 1, pp. 111–119, 2024, doi: 10.34010/komputa.v13i1.12639.

[9] Fadellia Azzahra, N. Suarna, and Y. Arie Wijaya, "Penerapan Algoritma Random Forest Dan Cross Validation Untuk Prediksi Data Stunting," *Kopertip J. Ilm. Manaj. Inform. dan Komput.*, vol. 8, no. 1, pp. 1–6, 2024, doi: 10.32485/kopertip.v8i1.238.

[10] D.- Husen, D.- Sandi, S.- Bumbungan, K.- -, and K.- -, "Analisis Prediksi Kebakaran Hutan dengan Menggunakan Algoritma Random Forest Classifier," *Nuansa Inform.*, vol. 16, no. 1, pp. 150–155, 2022, doi: 10.25134/nuansa.v16i1.5392.

[11] N. Maulidah, M. Maulidah, R. Supriyadi, H. Nalatissifa, S. Diantika, and A. Fauzi, "Prediksi_Kualitas_Air_Menggunakan_Metode_Random_Fo," *J. Khatulistiwa Inform.*, vol. 12, no. 1, pp. 1–6, 2024.

[12] S. P. Tamba and E. -, "Prediksi Penyakit Gagal Jantung Dengan Menggunakan Random Forest," *J. Sist. Inf. dan Ilmu Komput. Prima(JUSIKOM PRIMA)*, vol. 5, no. 2, pp. 176–181, 2022, doi: 10.34012/jurnalsisteminformasidanilmukomputer.v5i2.2445.

[13] A. Sharma, "Loan_Approval_Dataset," 2023. Accessed: Jul. 10, 2025. [Online]. Available: https://www.kaggle.com/datasets/architsharma01/loan-approval-prediction-dataset/data